JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR VOLUME 7, NUMBER 1 JANUARY, 1964

AVTA: A DEVICE FOR AUTOMATIC VOCAL TRANSACTION ANALYSIS[1]

LOUIS CASSOTTA, STANLEY FELDSTEIN, AND JOSEPH JAFFE[2]

THE WILLIAM ALANSON WHITE INSTITUTE AND COLUMBIA UNIVERSITY

The Automatic Vocal Transaction Analyzer was designed to recognize the pattern of certain variables in spontaneous vocal transactions. In addition, it records these variables directly in a machine-readable form and preserves their sequential relationships. This permits the immediate extraction of data by a digital computer. The AVTA system reliability has been shown to be equal to or better than that of a trained human operator in uncomplicated interaction. The superiority of the machine was demonstrated in complex interactions which tax the information processing abilities of the human observer.

A critical problem in the investigation of verbal behavior is that of extracting and processing large quantities of data. The problem is particularly acute when tape recorded spontaneous speech is studied by traditional techniques. The time and effort required to extract such attributes as pauses within speech, speech rate, or the interactional behavior of two or more speakers, has severely limited the scope of pertinent research. Thus, for example, one recent book is entirely devoted to the first 5 min of an interview (Pittenger, Hockett, and Danehy, 1960), while another is concerned with only two of an extended series of psychotherapy sessions (Gottschalk, 1961). Many individual experimental studies are similarly limited.

Several approaches have already been made to a solution of the data processing problem. Kasl and Mahl (1956) devised an instrument to measure the duration of silences within a fixed period of time. Chapple's (1949) Interaction Chronograph reliably describes an interview situation in terms of such interaction parameters as nods, gestures, verbal interruptions, silences, and others. While these devices have decreased the participation of the human operator in the extraction and processing of data, they have not completely eliminated him. A trained observer is still required to mediate between the behavior and the equipment. Exceptions to this requirement are the device Verzeano and Finesinger (1949) used to obtain certain interaction variables automatically, the instrument Starkweather (1960) designed to obtain an estimate of speech rate, and the equipment he and Hargreaves (1961) are developing to obtain vocal frequency and intensity. A limitation of all of the above equipment is that its output is not in a form amenable to immediate statistical evaluation by computer programs. Thus, further human intervention is needed either to perform the analyses or to prepare the data for computer processing. These necessary stages of human intervention are typical bottlenecks which constrain the scope of research programs in psychology.

The equipment described here, the Automatic Vocal Transaction Analyzer (AVTA), grew out of the needs of a research program now in progress. AVTA is intended to bridge the gap between tape recorded interviews and IBM punch cards. It has been designed to "listen" to tape recordings of dyadic conversa-

[1]Based upon a paper presented at the Eastern Psychological Association, Atlantic City, April, 1962. The work was supported in part by the following grants from the National Institute of Mental Health: Research Grants M-4548, M-4571, and MH-04571-03 to The William Alanson White Institute, and a General Research Support Grant to the College of Physicians and Surgeons, Columbia University. The consultative assistance of the IBM Corporation is gratefully acknowledged. A working model of the AVTA system was developed at the psycholinguistic laboratory of The William Alanson White Institute. The final design was completed and fabricated by Scientific Prototype Corporation, New York City. Reprints may be obtained from Louis Cassotta, Research Department, William Alanson White Institute, 12 East 86th St., New York 28, N.Y.

[2]Dept. of Psychiatry, Columbia University and also The William Alanson White Institute.

tion, recognize certain variables of interest, and automatically punch the data at regular intervals, and in machine-readable form.

AVTA System

Figure 1 is a partial block diagram of the AVTA unit. The unit is connected to the output of a two channel tape recorder. The interview is so recorded that the voice of each participant is primarily confined to one channel. Each of the channels operates in the following manner. The audio signal of the associated speaker is amplified and rectified. The rectified signal is then filtered, producing a dc signal proportional to the recorded voice. This signal is passed through a threshold level device. When the amplitude of the dc signal exceeds the pre-set threshold level, it actuates an amplifier driven relay. The threshold level is set high enough to exclude random noise, but low enough to include the voice of the speaker when speaking softly. Thus, the state of the output relay (de-energized or energized) indicates the presence or absence of verbal behavior on the part of the associated speaker. The amount of filtering (or time constant) and the threshold level desired for particular experimental purposes may be selected by a front panel control knob.

The outputs of these relays are used to actuate an IBM keypunch directly. System operation requires that the keypunch receive a command at regular intervals, to "inquire" about the state of each of the relays. The command is generated by an electronic timing device located in the AVTA chassis. (A front panel control knob permits the operator to set the punching rate at any interval from 1 per sec to 10 per sec.) At the end of each interval, the timer readies the keypunch to receive the data punching signals from the output relays. Thus, if speaker A is talking at the interval end, relay A will initiate a punch in a designated row of the current column of the IBM card. Simultaneously, relay B will cause the keypunch to record the presence of speech of speaker B in a second row of the same column. The card is then automatically advanced one column to await further instructions forthcoming at the end of the next interval. In this manner, the IBM cards receive a continuous record of the presence or absence of verbal behavior on the part of each participant in the interview.

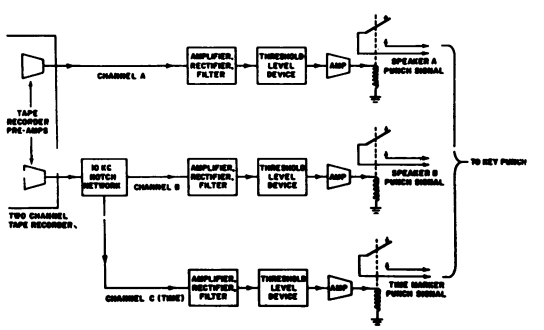

Fig. 1. A block diagram depicting the speaker and time interval channels of AVTA.

In order to provide a time axis, a 10 kc signal is superimposed upon one of the audio-tape channels for a duration of 1 sec at regular intervals during the initial recording. AVTA extracts the signal by passing it through the notch network shown in Fig. 1. The signal then actuates relay C as indicated in the drawing. Relay C is interconnected with a third designated row on the keypunch, thereby placing a time axis on the data cards.[3]

The system thus far described would not adequately handle the problem created by the spilling of one speaker's voice into the microphone of the other. This unintended recording of one speaker's voice in the other's channel can cause both output relays to be actuated when only one participant is speaking. The relative proximity of each speaker's voice to his own microphone provides a considerable degree of attenuation of the voice of the nonintended speaker. The remainder of the nonintended vocal signal is cancelled electronically. The cancellation network shown in Fig. 2 carries an equal amount of signal from its intended channel which is used to subtract, or cancel, the effect of the recording of one speaker in the other speaker's channel. The amount of cancellation signal required depends upon such factors as the distance of the speakers from each other. The requirement for any particular recording is determined empirically by adjusting a front panel control knob to the minimum setting required to eliminate the spilled signal. The use of this cancellation network permits flexibility in

[3]The state of each of the output relays is visually displayed by a lamp on the front panel. Thus, the lighting of a lamp signals an observer that its associated relay is energized. The presence and absence of the time marker signal is also indicated by a lamp.

the experimental setting as well as the use of a wider range of microphones than could otherwise be employed.

A Fortran program has been written which enables an IBM 7090 computer to extract and quantify many of the variables inherent in the card output of the AVTA system (Jaffe, Feldstein and Cassotta, in press). The variables utilized in this program are utterances, pauses, vocal time, latencies, and three types of simultaneous speech. They are operationally defined as follows.

Utterance. An utterance is a sequence of punches (not necessarily consecutive) which may appear in either speaker's row. It is terminated according to the following rules. A vocalization by one speaker during any silence of the other speaker is considered a termination of the latter's utterance. The utterance is considered terminated at the last punch in the speaker's row that precedes his silence. The utterance is also terminated if the speaker's resumption, after a period of silence, appears in the same column as the onset of speech of the other speaker. Again, an utterance is considered terminated if one speaker stops talking at the same time as the other speaker, so that the last punch (preceding a period of mutual silence) of each speaker appears in the same column.[4] An utterance of one speaker may be simultaneous with that of the other speaker, in which case it is classified in terms of one of the categories of simultaneous speech (see below).

Pause. A pause is one or more unpunched columns within an utterance. Reference to the definition of utterance termination indicates that only an interval of silence within an utterance will be considered a pause. If A's utterance is terminated, any intervening silence is considered a latency attributable to B.

Latency. A latency is a period of silence between the termination of one speaker's utterance and the onset of the other speaker's utterance. Only positive latencies are recorded. If B starts speaking before A has terminated, the latency is technically negative but is, instead, dealt with by a category of simultaneous speech.

Simultaneous Speech. Simultaneous speech consists of those parts of two utterances which are temporally concurrent. Three subtypes are distinguished.

(1) *Nonencompassed speech* is that which (a) starts later but persists longer than the concurrent speech, or (b) starts simultaneously but ends later than the concurrent speech.

(2) *Encompassed speech* is that which (a) starts later and ends earlier than the concurrent speech, or (b) starts later and ends simultaneously with the concurrent speech.

(3) *Synchronous utterances* are those which start and end simultaneously and are credited to both speakers.

Figure 3 depicts an IBM card which has been processed through AVTA and demonstrates representative types of interactions encountered.

The computer provides printouts for each of these variables in terms of their summed time for each time unit of the interview. In addition, the frequency of occurrence of each of the variables is computed and read out for each time unit of the interview. Finally, the computer produces histogram data for each of these variables plus a variety of descriptive statistics. The histogram interval may be set to any desired value. Typically, the system has been most frequently set to produce histogram data in .01 min intervals.

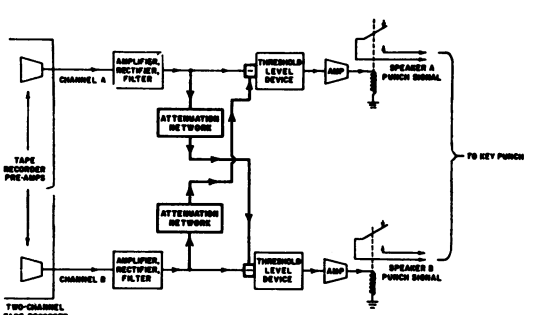

Fig. 2. A diagram of the placement of the non-intended signal cancellation network.

Reliability Studies

Two studies were conducted to evaluate the reliability of the AVTA system and to compare its performance to that of a highly trained operator. The first study was primarily designed to test the capacity of the AVTA system to detect low-level vocal signals reliably. For this purpose, the two participants

[4]This definition of an *utterance* is essentially the same as the definition of *utterance unit* given by Fries (1952, p. 23).

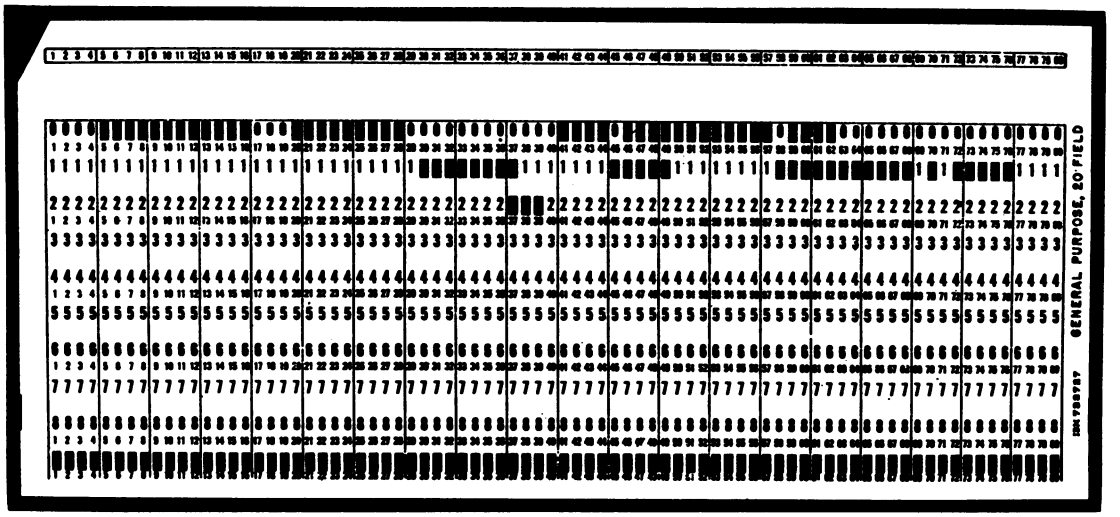

Fig. 3. A composite record demonstrating representative types of interactions encountered. Each column (col.) represents 0.3 sec in the record. Row 9 is punched continuously, as a check on the command operation. The time unit indication appears in row 2 in cols. 37-39. Speaker A's channel records in Row 0, speaker B's in Row 1. The first four columns indicate complete silence. Speaker A starts in col. 5, he pauses in cols. 17, 18, 19, then resumes in 20. He ceases in col. 28. Col. 29 is a latency. B begins speaking in col. 30 and continues to col. 37. After a latency in cols. 38-40, A speaks again. During A's pause in col. 45, B begins speaking. However, A continues in col. 46, and for the next four columns A and B speak simultaneously. B stops speaking in col. 50. The interaction is repeated in col. 58, when B commences during a pause in A's vocalization. This time, however, A stops after four columns of simultaneous speech.

were instructed to speak in relatively soft tones. The resulting 20-min recording was a rather slowly paced conversation in which the incidence of simultaneous speech was comparatively rare. In the second study, the subjects were encouraged to speak freely in varying degrees of loudness. Previous work, including the first study, had indicated that normal free conversation did not produce sufficient simultaneous speech to permit generation of reliability figures in these categories. Therefore, in this second study, the subjects were instructed to interrupt each other frequently. In all other respects, the conversations typified free speech in that no artificial structure was employed. Rather, the subjects were given an initial topic to discuss and the conversation was permitted to flow freely from that point on. In the second study, the recording was 30 min long, employing four speakers (two at a time).

The tape recordings were each analyzed by the AVTA system on three different occasions, thereby generating three sets of data cards for each study. Product moment correlation coefficients between paired occasions for all variables were computed. The reliability estimate for each variable is the average of the three correlation coefficients for that study. These data for studies I and II are presented in Tables I and II respectively. In study I, simultaneous speech occurred too infrequently to permit computation of reliability data in these categories. While the second study includes reliability figures for encompassed and nonencompassed speech, the incidence of synchronous speech was still too infrequent to permit generation of reliability estimates for that category.

Table I

Reliability Estimates of AVTA and a Human Operator Based upon Successive 1-min Speech Samples for Reliability Study I

Variable	*AVTA*	*Human Operator*
Frequency of Utterances	.989	.877
Total Vocal Time	.995	.895
Sum of Utterance Lengths	.993	.979
Sum of Latency Durations	.916	.899
Sum of Pause Lengths	.989	.873
Frequency Distribution of Pause Lengths	.997	.987
Frequency Distribution of Utterance Lengths	.992	.888

Table II

Reliability Estimates of AVTA and a Human Operator Based upon Successive 2-min Speech Samples for Reliability Study II

Variable	Reliability Estimates AVTA	Reliability Estimates Human Operator
Sum of Encompassed Speech	.907	.624
Sum of Nonencompassed Speech	.951	.643
Frequency of Encompassed Speech	.953	.645
Frequency of Nonencompassed Speech	.882	.595
Frequency of Utterances	.974	.721
Total Vocal Time	.999	.885
Sum of Utterance Lengths	.999	.894
Sum of Latency Durations	.967	.679
Sum of Pause Lengths	.988	.848
Frequency Distribution of Pause Lengths	.987	.939
Frequency Distribution of Nonencompassed Speech Lengths	.932	.850
Frequency Distribution of Encompassed Speech Lengths	.968	.925
Frequency Distribution of Utterance Lengths	.969	.945
Frequency Distribution of Latency Durations	.977	.986

A trained human operator was required to perform the same function by pressing the appropriate switches as each subject spoke in the interview. (These switches were wired to replace the AVTA output relays, thereby permitting the operator to simulate the AVTA system.) The averaged correlation coefficients of the three runs for the human operator were then computed as reliability estimates of human performance. These are also presented in Tables I and II. The data are based upon comparisons of successive 1-min intervals for the summed time of the variables and for the frequency of the speech units. For study II, the data is based upon successive 2-min intervals. The histogram data reliability was computed by correlating the incidence of each variable for each successive interval on a total interview basis.

Examination of these data shows clearly that the reliability of AVTA is consistently higher than that of the trained operator. In addition, comparison of study I and study II data indicates that the trained operator's performance is seriously impaired when the recording contains a high incidence of simultaneous speech, whereas the reliability of AVTA is not noticeably affected by this condition. Saslow and Matarazzo (1959) also found, in a study utilizing the Interaction Chronograph, that interscorer reliability was lowest when the incidence of simultaneous speech increased. The advantages of automated electronic processing becomes apparent when task difficulty begins to tax the information processing abilities of human scorers.

The reliability of AVTA is relatively independent of the quality of the tape recording. Errors introduced by sounds other than speech will not decrease the reliability of the system in that AVTA will consistently record these. However, since the purpose of the system is to record the temporal patterns of speech, extraneous noise will decrease the validity of system operation. Thus, system validity is a function of the conditions of recording. If the interview is conducted within a room insulated from outside noise, or if noise cancellation microphones are utilized, the incidence of external noise can be reduced to a negligible amount. While satisfactory recording conditions can be readily achieved, precautions must be taken to insure that such conditions are met. An operator, listening briefly to the recording while watching the AVTA output, can readily assess the extent to which irrelevant noise impairs the validity of the AVTA data.

The development of AVTA was, of course, based upon the assumption that the variables it was designed to extract are meaningfully related to certain underlying psychological phenomena. The assumption is partially supported by the work of Chapple (1954), Feldstein and Jaffe (1962), Goldman-Eisler (1958), Hargreaves (1960), Jaffe, Feldstein, and Cassotta (1962), Saslow and Matarazzo (1959), and others. AVTA makes feasible the exploration and evaluation of these variables more rapidly, with less effort, greater precision, and in terms of larger samples of data than has hitherto been possible.

REFERENCES

Chapple, E. D. The interaction chronograph: Its evolution and present application. *Personnel*, 1949, **25**, 295-307.

Chapple, E. D., Chapple, Martha F., and Repp, Judith A. Behavioral definitions of personality and temperament characteristics. *Human Organiz.*, 1954, **13**, 34-39.

Feldstein, S. and Jaffe, J. A note about speech disturbances and vocabulary diversity. *J. Communication*, 1962, **12**, 166-170.

Fries, C. C. *The structure of English*. New York: Harcourt, Brace & World, 1952.

Goldman-Eisler, Frieda. Speech analysis and mental process. *Lang. and Speech*, 1958, **1**, 59-75.

Gottschalk, L. A. (Ed.). *Comparative psycholinguistic analysis of two psychotherapeutic interviews*. New York: Int. Univer. Press, 1961.

Hargreaves, W. A. A model for speech unit duration. *Lang. and Speech*, 1960, **3**, 164-173.

Hargreaves, W. A. and Starkweather, J. A. Vocal behavior: An illustrative case study. Paper read at West. Psychol. Assn., Seattle, June, 1961.

Jaffe, J., Feldstein, S., and Cassotta, L. An IBM 7090 program for analyzing vocal parameters of dyadic interaction. *Behav. Sci.*, in press (Abstract).

Jaffe, J., Feldstein, S., and Cassotta, L. A model for the temporal description of vocal interaction. Unpublished manuscript, The William Alanson White Institute, New York City, 1962.

Kasl, S. V. and Mahl, G. F. A simple device for obtaining certain verbal activity measures during interviews. *J. abnorm. soc. Psychol.*, 1956, **53**, 388-390.

Mahl, G. F. Exploring emotional states by content analysis. In I. Pool (Ed.), *Trends in content analysis*. Urbana: Univ. Illinois Press, 1960, p. 83-130.

Saslow, G. and Matarazzo, J. D. A technique for studying changes in interview behavior. In E. A. Rubenstein and M. B. Parloff (Eds.), *Research in psychotherapy*. Washington, D.C.: Amer. Psychol. Assn., 1959, p. 125-159.

Starkweather, J. A. A speech rate meter for vocal behavior analysis. *J. exp. Anal. Behav.*, 1960, **3**, 111-114.

Verzeano, M. and Finesinger, J. E. An automatic analyzer for the study of speech in interaction and in free association. *Science*, 1949, **110**, 45-46.

Received September 26, 1963